

BioMedBridges

E-Infrastructure support for the life sciences:
Preparing for the data deluge

Challenges of big data
Aims of the workshop

Rafael Jimenez

ELIXIR CTO

15 May, 2014

Second BioMedBridges AGM

10-11 March 2014
Florence, Italy

**e-infrastructure advisory board meeting
with BMS RI technical representatives**

FEBRUARY 27TH - MARCH 5TH 2010

Economist.com

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT



Economist

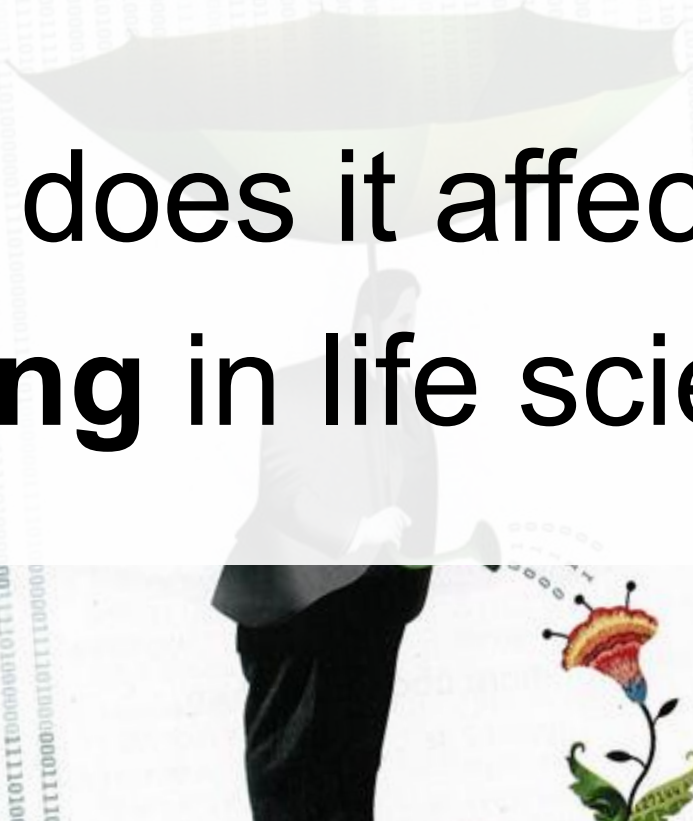
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs

FEBRUARY 27TH - MARCH 5TH 2010

Economist.com

The data deluge

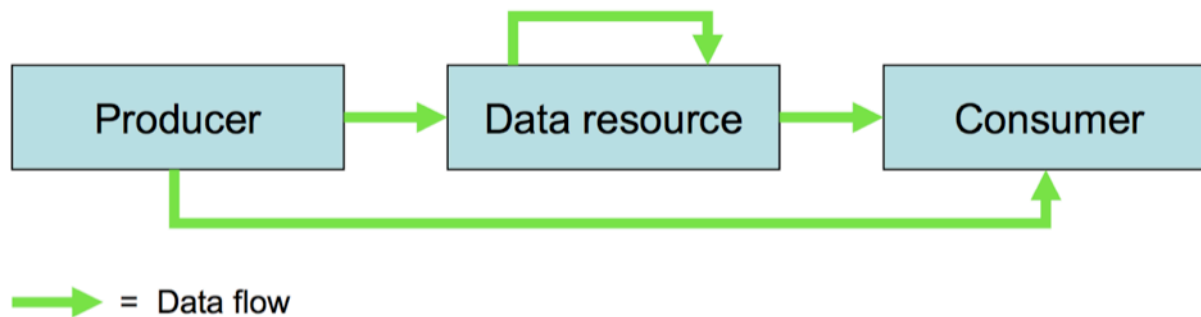
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT



How does it affect data sharing in life sciences?

Large-scale data sharing in the life sciences

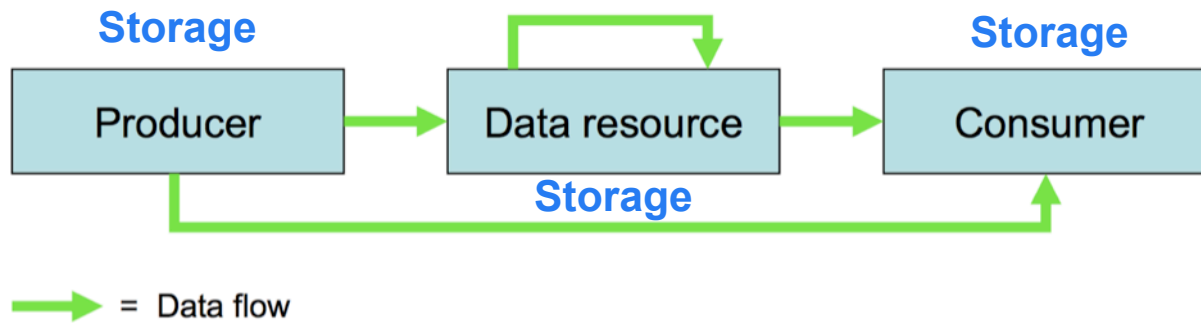
Sharing involves a **producer**¹ who is the source of what is to be shared (often its creator), and a **consumer** (sometimes called a customer, user or recipient). In some cases one or more **data resources** which store and/or make data available may lie between the producer and the consumer.



Sharing can happen where consumers are sent information, by a **push** from the producer or data resource, or the consumer can seek the information to be shared, and **pull** it to him or herself.

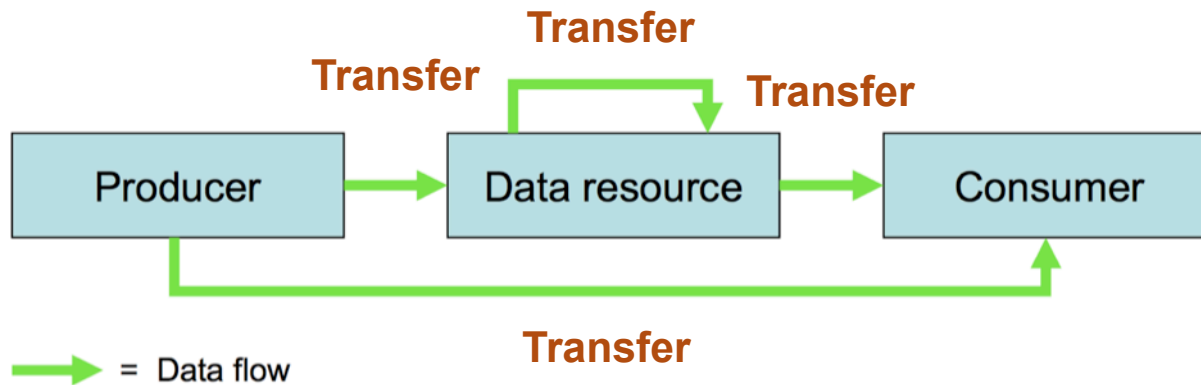
How does big data affect data sharing?

Storage



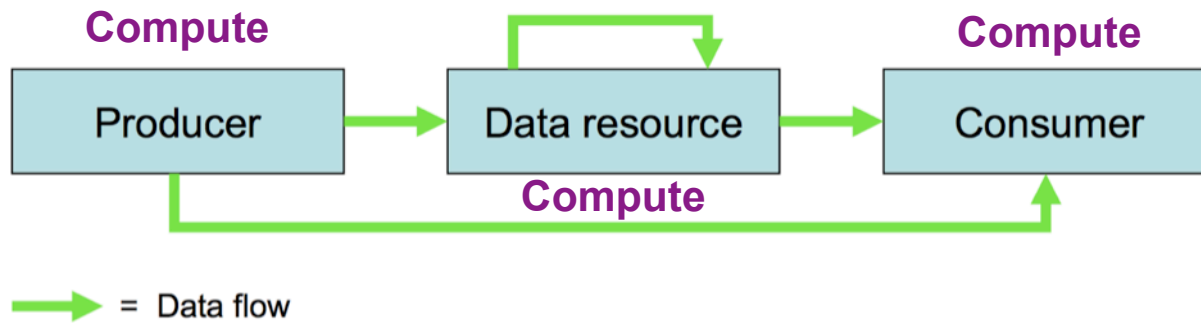
How does big data affect data sharing?

Transfer



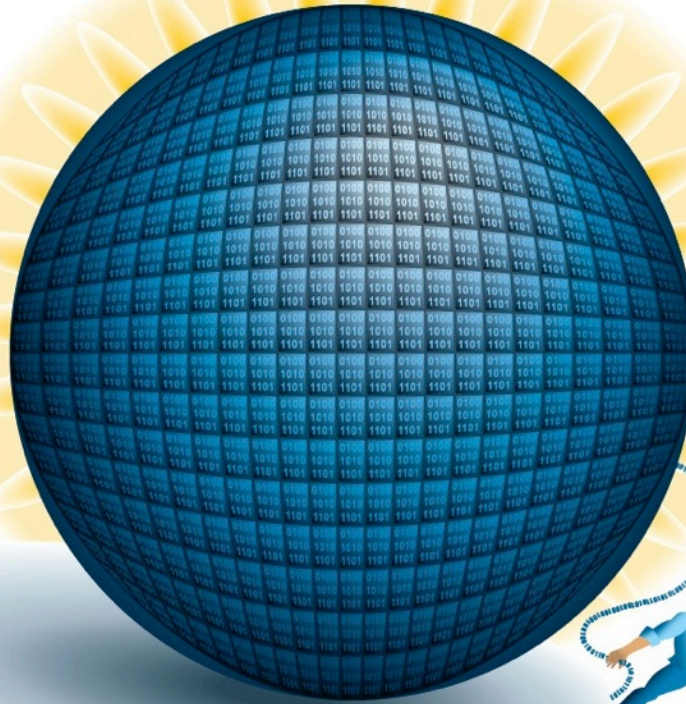
How does big data affect data sharing?

Compute



DATA MANAGED
WILL
**INCREASE BY
50
TIMES**

IT
PROFESSIONALS
WILL
**INCREASE BY
1.5
TIMES**



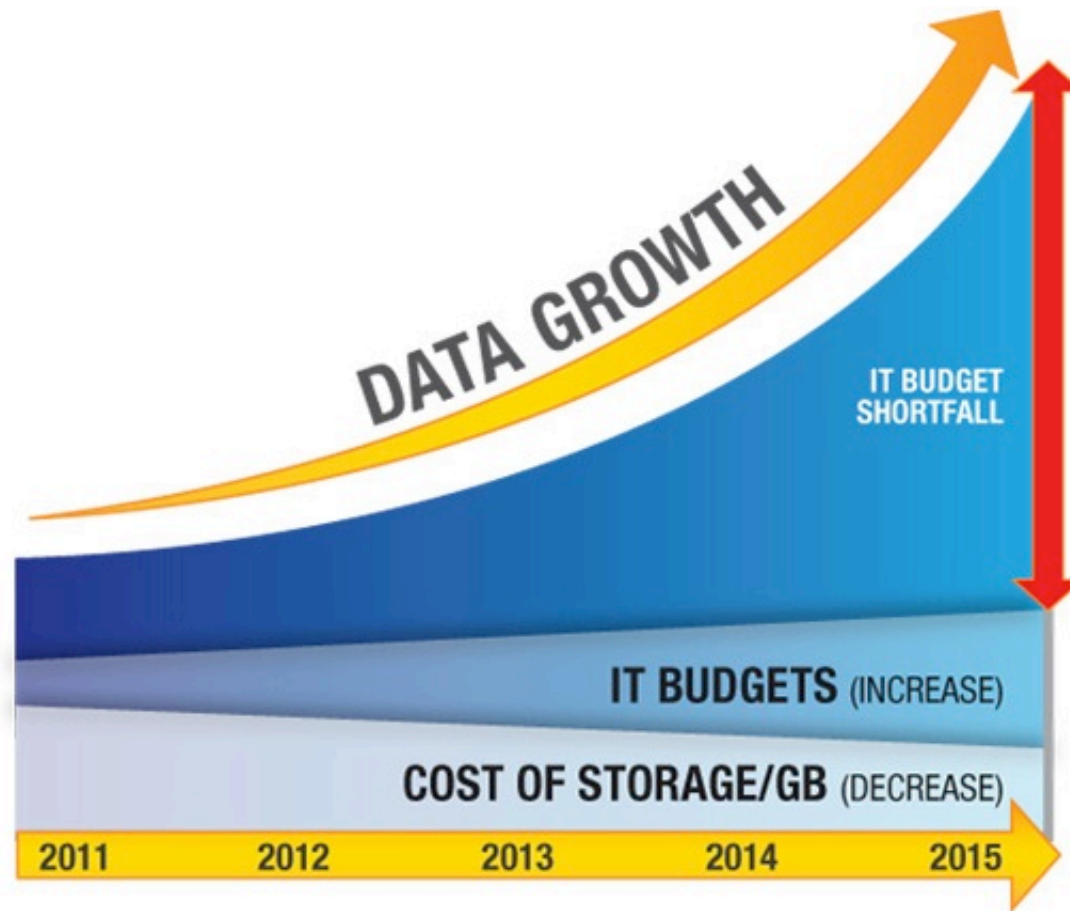
DATA
MANAGED IT PROFESSIONALS
2011

DATA
MANAGED IT PROFESSIONALS
2020

THE
2011
IDC
**DIGITAL
UNIVERSE** STUDY
sponsored by EMC

Data growth

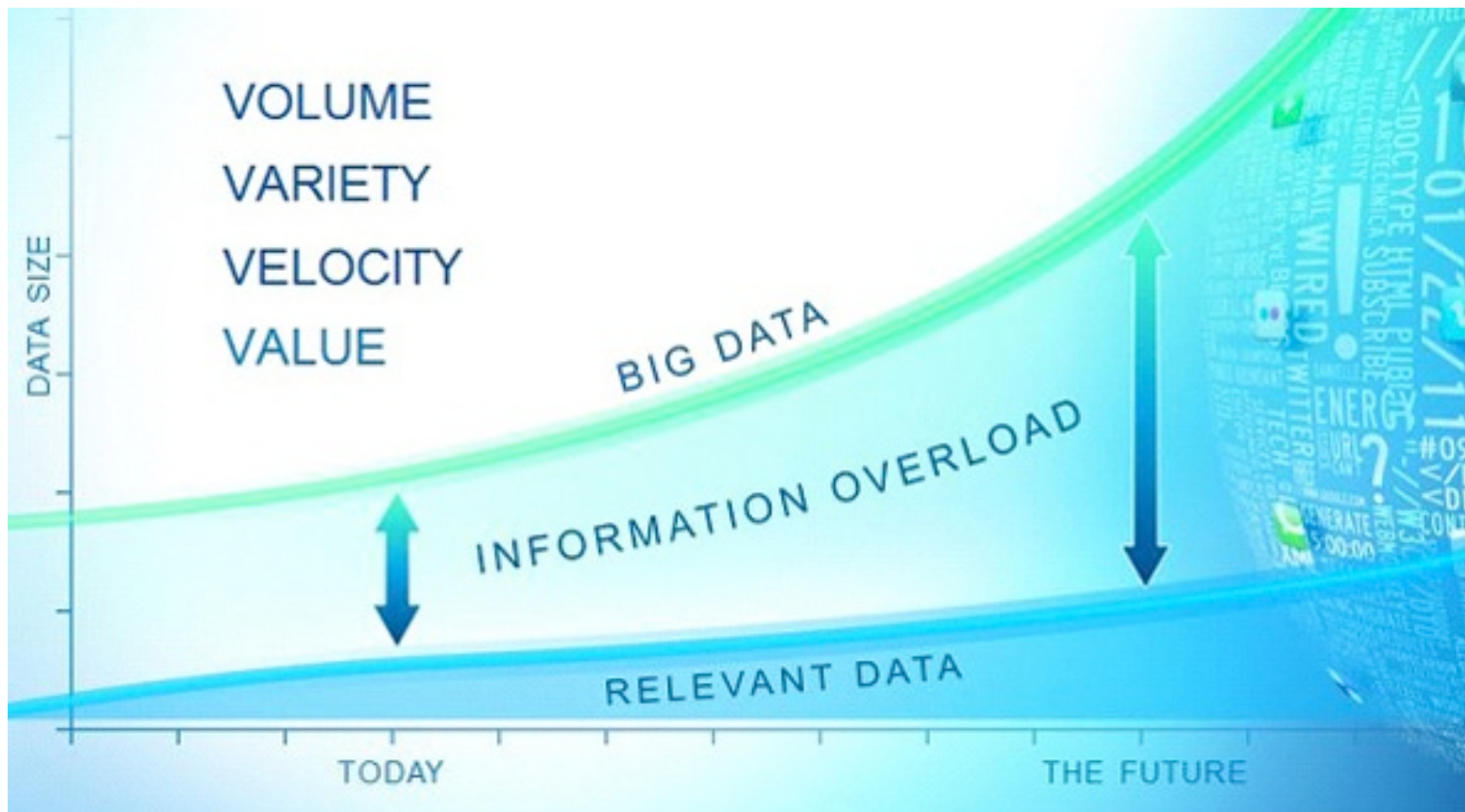
how to reduce the IT budget shortfall?



<http://www.eweek.com/>

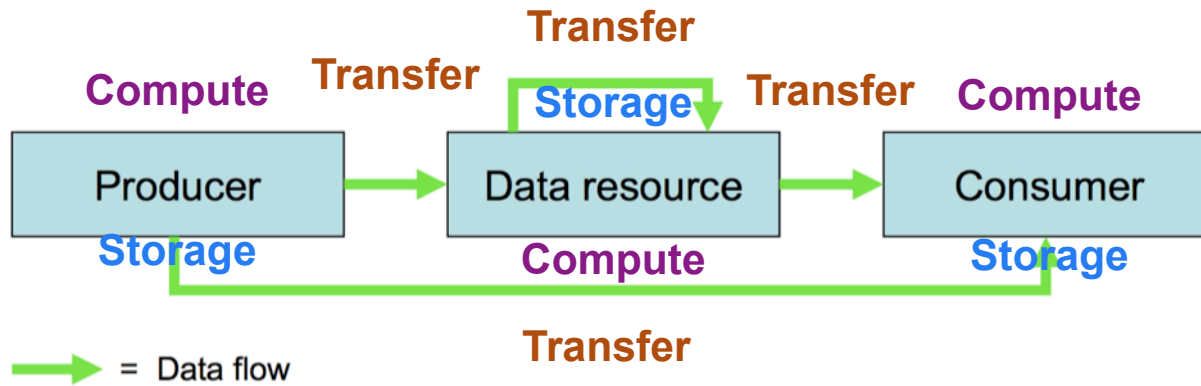
Data growth

What data is relevant?



Problems of big data

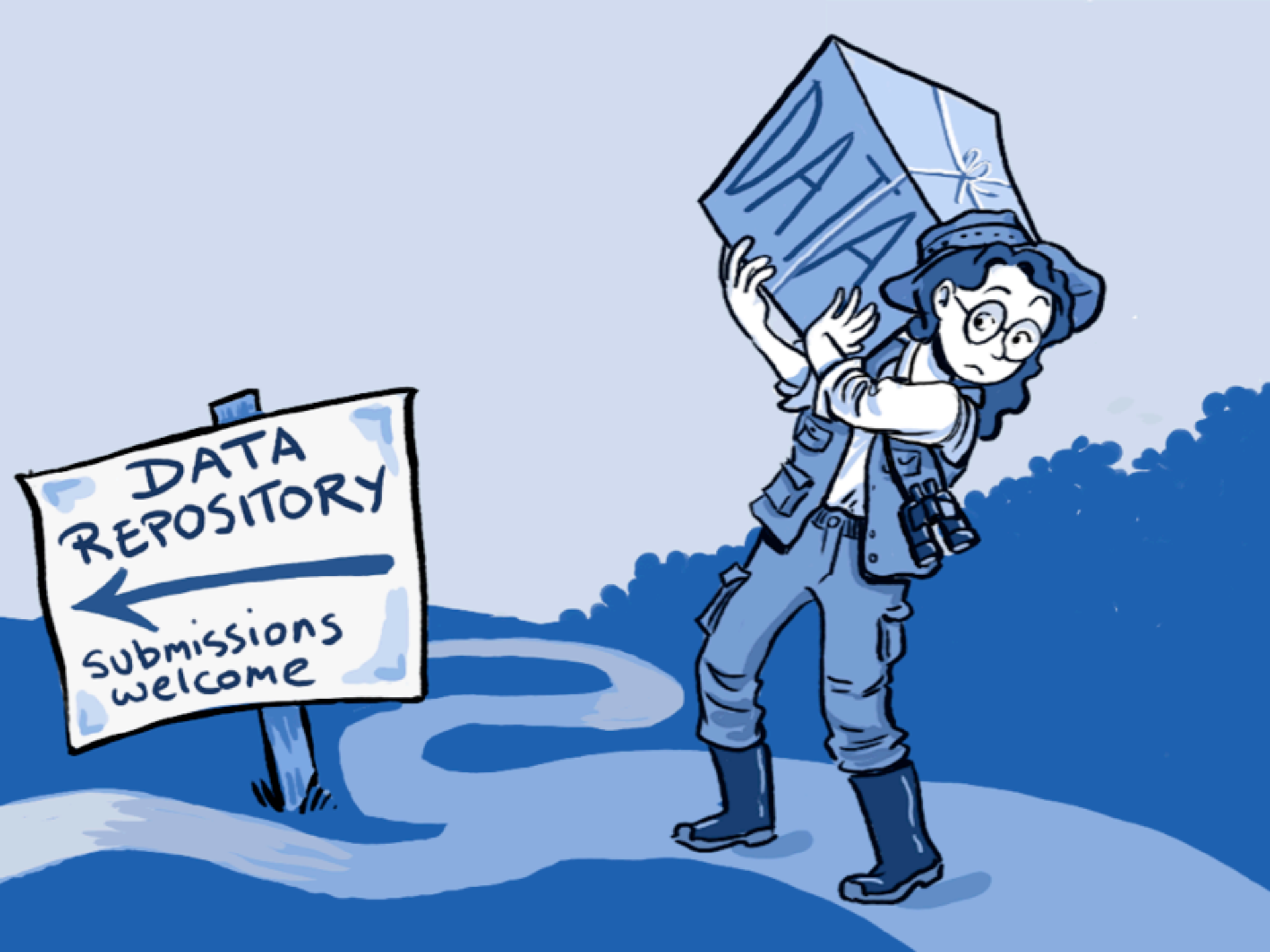
Storage Compute Transfer



What How Where



<http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002552>

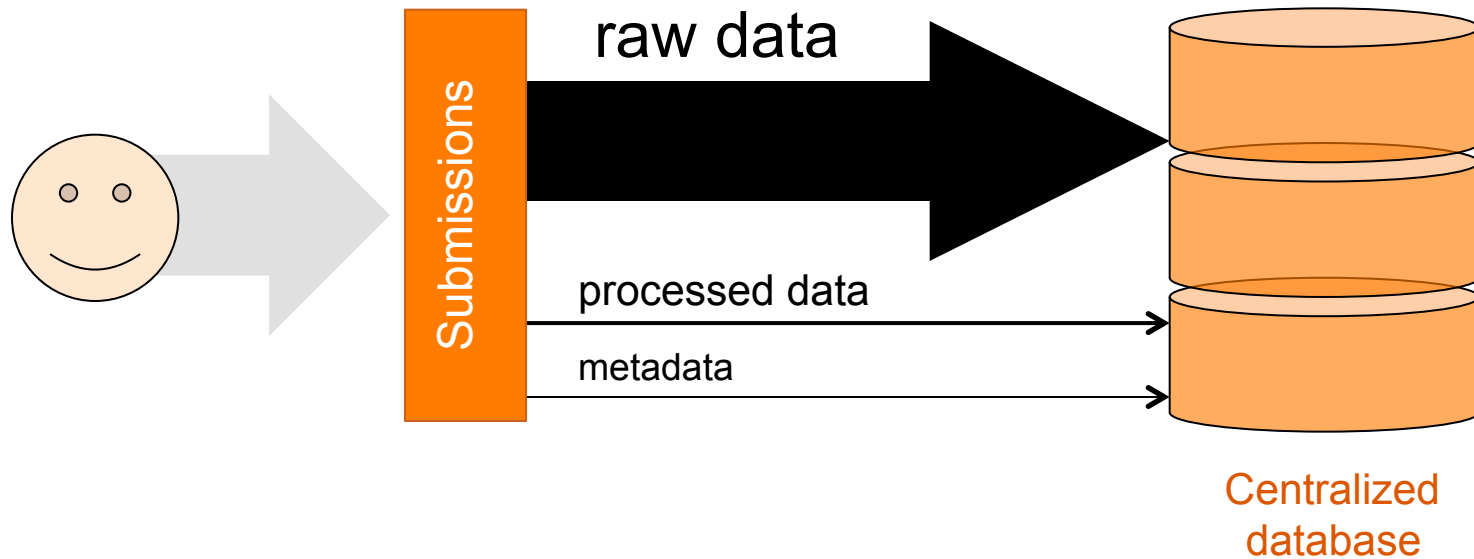


DATA
REPOSITORY



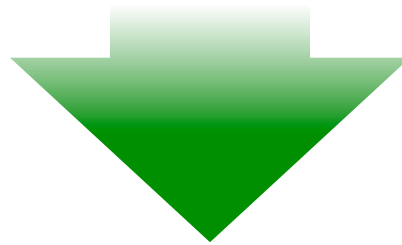
submissions
welcome

Data submission



Data sharing

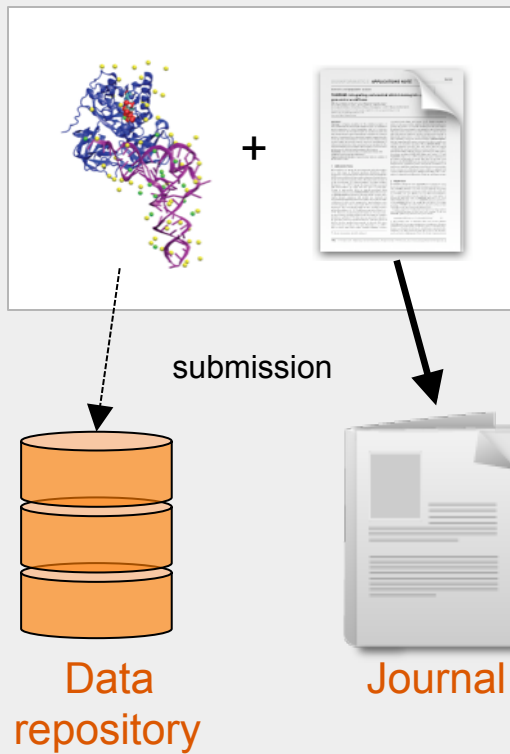
The casual approach 'data on my disk and available to anyone who requests it'



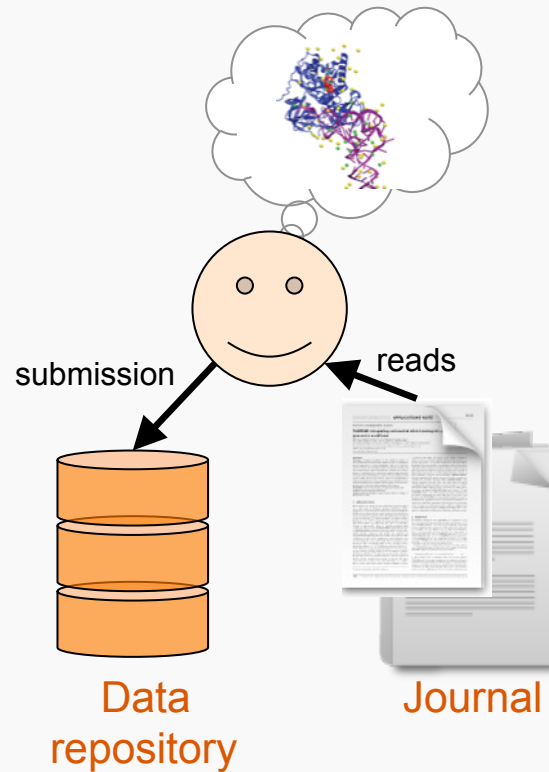
Submission to data repositories

Data submissions

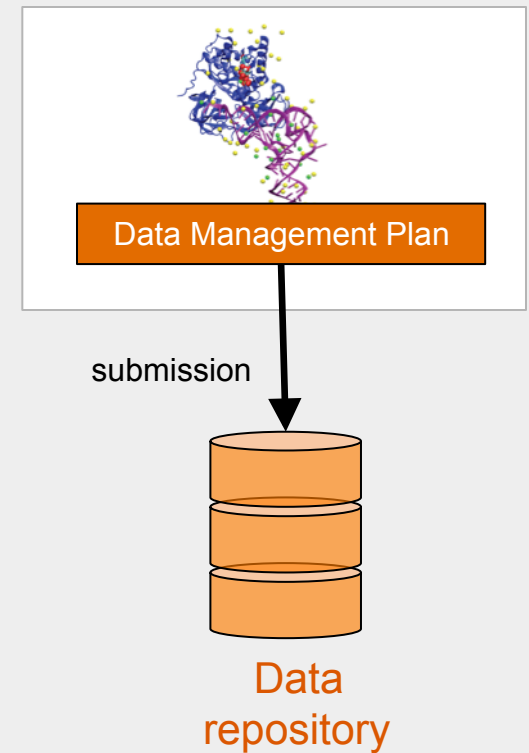
Journal request



Curator



Data management



Data sharing

Will big data affect data deposition?

The casual approach 'data on my disk and available to anyone who requests it'

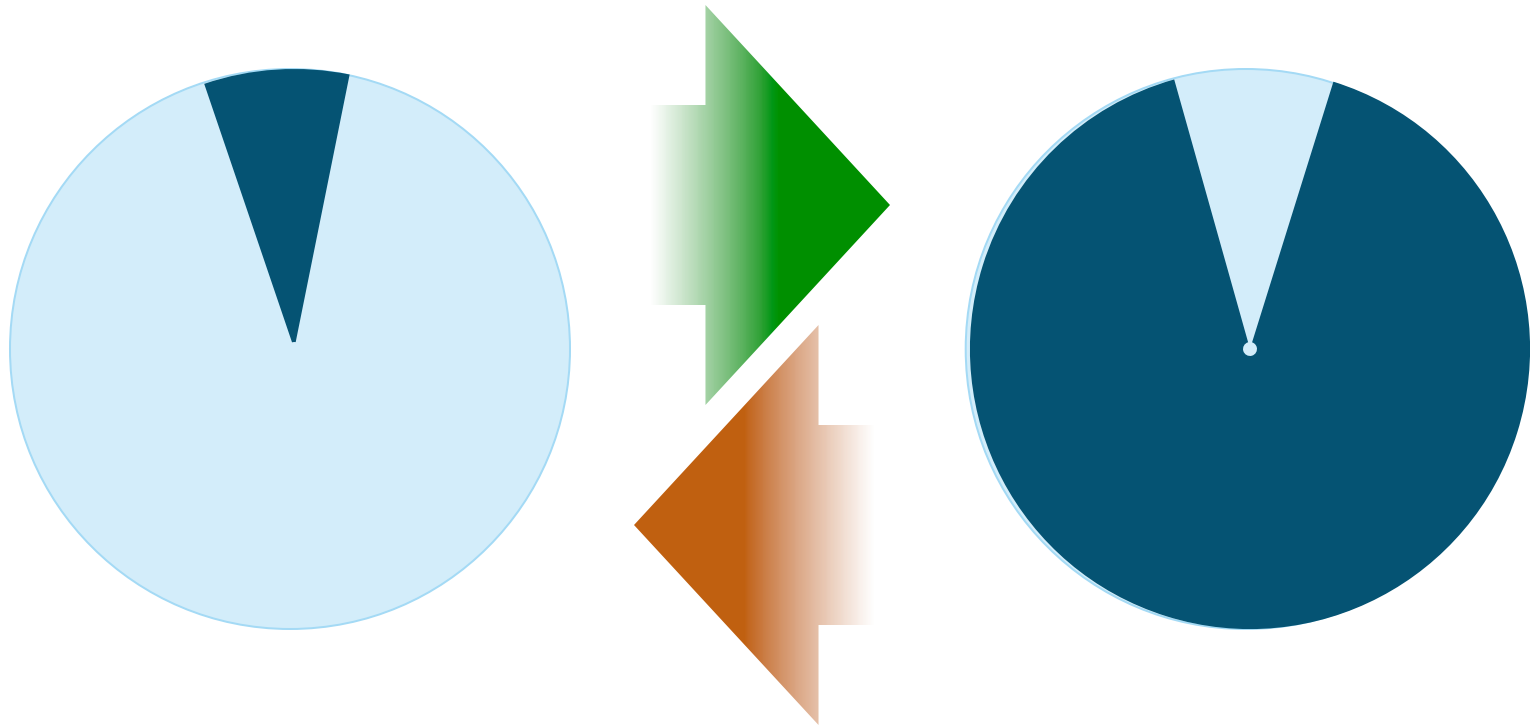


Submission to data repositories

Data submissions

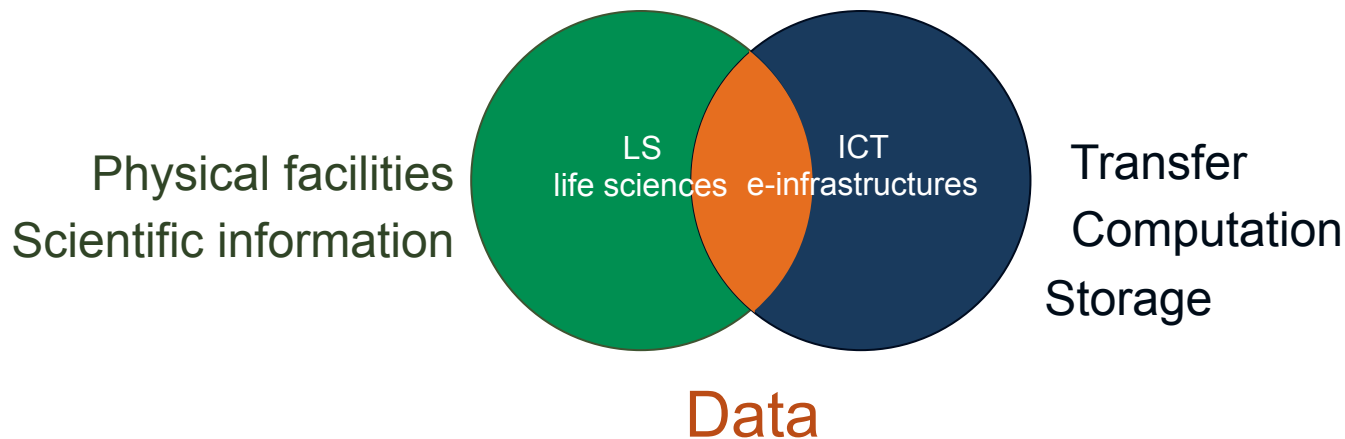
How much data?

How much available data?



Knowledge exchange workshop

- Discussion of big data challenges in life sciences
 - Focus on few representative domains
 - Looking 5 years ahead
- Jointly identify potential solutions to our problems

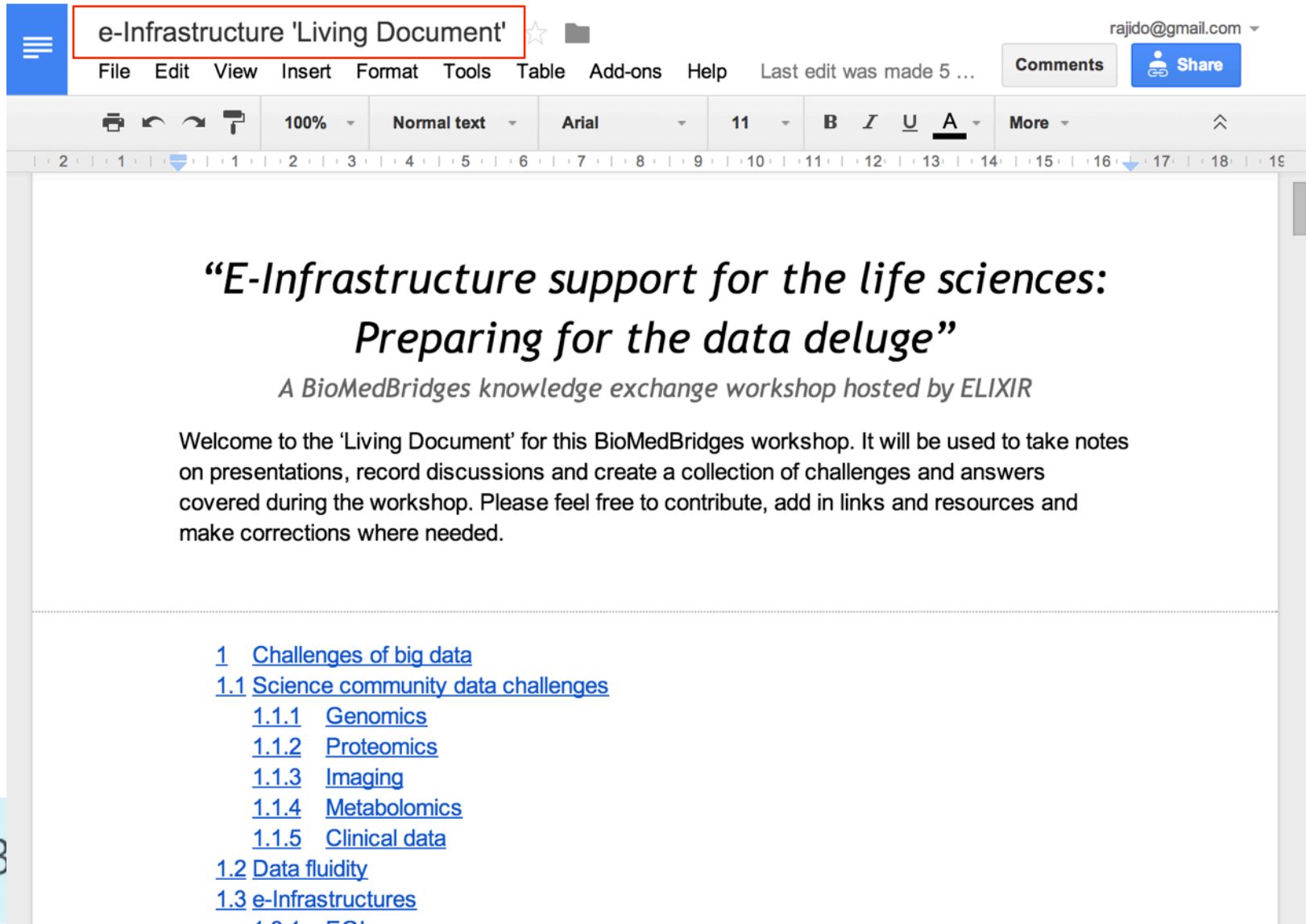


Data challenges of different science communities

Day 1: Thursday, 15 May		
12:00	Arrival and lunch	
13:00	Introductions	Tom Hancocks/Stephanie Suhr
13:10	Challenges of big data/aims of the workshop	Rafael Jimenez
13:30	<i>Flash presentations (5 minutes each): Data challenges of different science communities</i>	
	• Genomics	Pieter Neerincx (UMCG, BBMRI)
	• Proteomics	Henning Hermjakob (EMBL-EBI, ELIXIR)
	• Imaging	Jason Swedlow (U Dundee, Euro-BioImaging)
	• Metabolomics	Natalie Stanford (U Manchester, ISBE)
	• Clinical data	Jan-Willem Boiten (CTMM, EATRIS)
14:00	Data fluidity	Guy Cochrane (EMBL-EBI)
14:20	<i>Flash presentations (5 minutes each): e-Infrastructures</i>	
	• EGI	Tiziana Ferrari (EGI)
	• EUDAT	Per Öster (CSC)
	• GÉANT	Richard Hughes-Jones (GÉANT)
	• PRACE	Sergio Bernardi (PRACE)
	• CERN/LHC	Alberto di Meglio (CERN)
15:00	Round-up of challenges	Rafael Jimenez
15:10	Open discussion/Questions and Answers	
15:30	Break	
15:45	<i>Group sessions: Each group will work through a science community use case (Part 1)</i>	
16:15	Report from group sessions	
16:45	<i>Group sessions: Each group will work through a science community use case (Part 2)</i>	
17:30	Report from group sessions and discussion	

Questions to the living document
<http://tinyurl.com/bmb-ddw>

http://tinyurl.com/ienfra



The screenshot shows a web browser window with a TinyMCE editor. The document title is "e-Infrastructure 'Living Document'". The menu bar includes File, Edit, View, Insert, Format, Tools, Table, Add-ons, and Help. The status bar indicates "Last edit was made 5 ...". The toolbar shows various editing tools like undo, redo, bold, italic, underline, and text color. The main content area contains the following text:

***“E-Infrastructure support for the life sciences:
Preparing for the data deluge”***

A BioMedBridges knowledge exchange workshop hosted by ELIXIR

Welcome to the 'Living Document' for this BioMedBridges workshop. It will be used to take notes on presentations, record discussions and create a collection of challenges and answers covered during the workshop. Please feel free to contribute, add in links and resources and make corrections where needed.

[1 Challenges of big data](#)

[1.1 Science community data challenges](#)

- [1.1.1 Genomics](#)
- [1.1.2 Proteomics](#)
- [1.1.3 Imaging](#)
- [1.1.4 Metabolomics](#)
- [1.1.5 Clinical data](#)

[1.2 Data fluidity](#)

[1.3 e-Infrastructures](#)

[1.3.1 ESI](#)



e-infrastructures

Day 1: Thursday, 15 May		
12:00	Arrival and lunch	
13:00	Introductions	Tom Hancocks/Stephanie Suhr
13:10	Challenges of big data/aims of the workshop	Rafael Jimenez
13:30	<i>Flash presentations (5 minutes each): Data challenges of different science communities</i>	
	• Genomics	Pieter Neerincx (UMCG, BBMRI)
	• Proteomics	Henning Hermjakob (EMBL-EBI, ELIXIR)
	• Imaging	Jason Swedlow (U Dundee, Euro-BioImaging)
	• Metabolomics	Natalie Stanford (U Manchester, ISBE)
	• Clinical data	Jan-Willem Boiten (CTMM, EATRIS)
14:00	Data fluidity	Guy Cochrane (EMBL-EBI)
14:20	<i>Flash presentations (5 minutes each): e-Infrastructures</i>	
	• EGI	Tiziana Ferrari (EGI)
	• EUDAT	Per Öster (CSC)
	• GÉANT	Richard Hughes-Jones (DANTE)
	• PRACE	Sergio Bernardi (PRACE)
	• CERN/LHC	Alberto di Meglio (CERN)
15:00	Round-up of challenges	Rafael Jimenez
15:10	Open discussion/Questions and Answers	
15:30	Break	
15:45	<i>Group sessions: Each group will work through a science community use case (Part 1)</i>	
16:15	Report from group sessions	
16:45	<i>Group sessions: Each group will work through a science community use case (Part 2)</i>	
17:30	Report from group sessions and discussion	

Questions to the
 living document
<http://tinyurl.com/bmb-ddw>

Open discussion

Day 1: Thursday, 15 May		
12:00	Arrival and lunch	
13:00	Introductions	Tom Hancocks/Stephanie Suhr
13:10	Challenges of big data/aims of the workshop	Rafael Jimenez
13:30	<i>Flash presentations (5 minutes each): Data challenges of different science communities</i>	
	• Genomics	Pieter Neerincx (UMCG, BBMRI)
	• Proteomics	Henning Hermjakob (EMBL-EBI, ELIXIR)
	• Imaging	Jason Swedlow (U Dundee, Euro-BioImaging)
	• Metabolomics	Natalie Stanford (U Manchester, ISBE)
	• Clinical data	Jan-Willem Boiten (CTMM, EATRIS)
14:00	Data fluidity	Guy Cochrane (EMBL-EBI)
14:20	<i>Flash presentations (5 minutes each): e-Infrastructures</i>	
	• EGI	Tiziana Ferrari (EGI)
	• EUDAT	Per Öster (CSC)
	• GÉANT	Richard Hughes-Jones (DANTE)
	• PRACE	Sergio Bernardi (PRACE)
	• CERN/LHC	Alberto di Meglio (CERN)
15:00	Round-up of challenges	Rafael Jimenez
15:10	Open discussion/Questions and Answers	
15:30	Break	
15:45	<i>Group sessions: Each group will work through a science community use case (Part 1)</i>	
16:15	Report from group sessions	
16:45	<i>Group sessions: Each group will work through a science community use case (Part 2)</i>	
17:30	Report from group sessions and discussion	

Group sessions

Day 1: Thursday, 15 May		
12:00	Arrival and lunch	
13:00	Introductions	Tom Hancocks/Stephanie Suhr
13:10	Challenges of big data/aims of the workshop	Rafael Jimenez
13:30	<i>Flash presentations (5 minutes each): Data challenges of different science communities</i>	
	• Genomics	Pieter Neerincx (UMCG, BBMRI)
	• Proteomics	Henning Hermjakob (EMBL-EBI, ELIXIR)
	• Imaging	Jason Swedlow (U Dundee, Euro-BioImaging)
	• Metabolomics	Natalie Stanford (U Manchester, ISBE)
	• Clinical data	Jan-Willem Boiten (CTMM, EATRIS)
14:00	Data fluidity	Guy Cochrane (EMBL-EBI)
14:20	<i>Flash presentations (5 minutes each): e-Infrastructures</i>	
	• EGI	Tiziana Ferrari (EGI)
	• EUDAT	Per Öster (CSC)
	• GÉANT	Richard Hughes-Jones (DANTE)
	• PRACE	Sergio Bernardi (PRACE)
	• CERN/LHC	Alberto di Meglio (CERN)
15:00	Round-up of challenges	Rafael Jimenez
15:10	Open discussion/Questions and Answers	
15:30	Break	
15:45	<i>Group sessions: Each group will work through a science community use case (Part 1)</i>	
16:15	Report from group sessions	
16:45	<i>Group sessions: Each group will work through a science community use case (Part 2)</i>	
17:30	Report from group sessions and discussion	

Dinner

ca. 18:30	End of day 1	
19:00	Dinner at the Red Lion, Hinxton (hosted by ELIXIR)	
Day 2: Friday, 16 May		
09:00	Solutions for big data in other science communities	
	<ul style="list-style-type: none">• Earth satellite data	Wolfgang Lengert (ESA)
	<ul style="list-style-type: none">• Radio astronomy data	Arpad Szomoru (JIVE)
09:30	<i>Group sessions: Blue sky solutions for big data</i>	
10:30	Report from group sessions and discussion	
11:00	Break	
11:30	<i>Group sessions: Practicalities and actions needed to implement suggested solutions</i>	
12:30	Report from group sessions and discussion	
13:00	Lunch	
14:00	Closing discussion, roundup of challenges and solutions	Chair: Rafael Jimenez
15:00	End of workshop	

Summary of yesterday's discussion

ca. 18:30	End of day 1	
19:00	Dinner at the Red Lion, Hinxton (hosted by ELIXIR)	
Day 2: Friday, 16 May		
<hr/> <hr/>		
09:00	Solutions for big data in other science communities	
	<ul style="list-style-type: none">• Earth satellite data	Wolfgang Lengert (ESA)
	<ul style="list-style-type: none">• Radio astronomy data	Arpad Szomoru (JIVE)
09:30	<i>Group sessions: Blue sky solutions for big data</i>	
10:30	Report from group sessions and discussion	
11:00	Break	
11:30	<i>Group sessions: Practicalities and actions needed to implement suggested solutions</i>	
12:30	Report from group sessions and discussion	
13:00	Lunch	
14:00	Closing discussion, roundup of challenges and solutions	Chair: Rafael Jimenez
15:00	End of workshop	

Solutions for big data in other science communities

ca. 18:30	End of day 1	
19:00	Dinner at the Red Lion, Hinxton (hosted by ELIXIR)	
Day 2: Friday, 16 May		
09:00	Solutions for big data in other science communities	
	<ul style="list-style-type: none">• Earth satellite data	Wolfgang Lengert (ESA)
	<ul style="list-style-type: none">• Radio astronomy data	Arpad Szomoru (JIVE)
09:30	<i>Group sessions: Blue sky solutions for big data</i>	
10:30	Report from group sessions and discussion	
11:00	Break	
11:30	<i>Group sessions: Practicalities and actions needed to implement suggested solutions</i>	
12:30	Report from group sessions and discussion	
13:00	Lunch	
14:00	Closing discussion, roundup of challenges and solutions	Chair: Rafael Jimenez
15:00	End of workshop	

Group sessions

ca. 18:30	End of day 1	
19:00	Dinner at the Red Lion, Hinxton (hosted by ELIXIR)	
Day 2: Friday, 16 May		
09:00	Solutions for big data in other science communities	
	<ul style="list-style-type: none">• Earth satellite data	Wolfgang Lengert (ESA)
	<ul style="list-style-type: none">• Radio astronomy data	Arpad Szomoru (JIVE)
09:30	<i>Group sessions:</i> Blue sky solutions for big data	
10:30	Report from group sessions and discussion	
11:00	Break	
11:30	<i>Group sessions:</i> Practicalities and actions needed to implement suggested solutions	
12:30	Report from group sessions and discussion	
13:00	Lunch	
14:00	Closing discussion, roundup of challenges and solutions	Chair: Rafael Jimenez
15:00	End of workshop	

The goodness of Internet and TV. Times 100.

Gigabit speeds Crystal clear HD Cutting-edge DVR Your new remote All your content 1TB cloud storage Powerful Wi-Fi

Google fiber

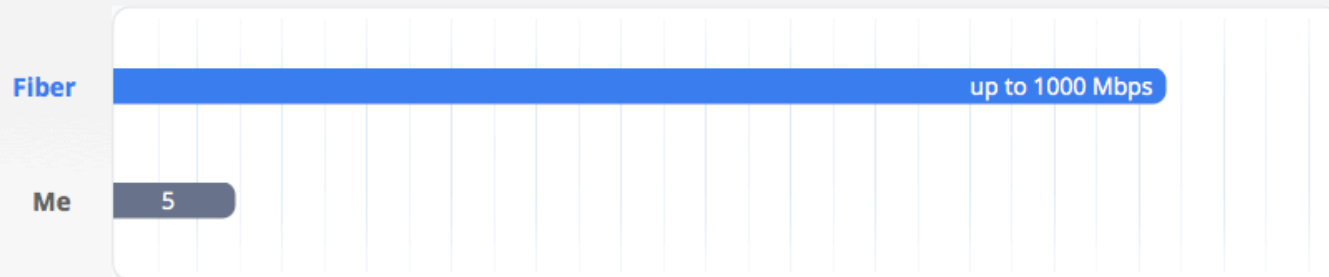
CITIES & PLANS ▾

SUPPORT

CHECK ADDRESS

rajido@gmail.com ▾

At up to 1,000 Mbps, Google Fiber is 100 times faster than today's basic broadband, allowing you to get what you want instantaneously. You no longer have to wait on things buffering; everything will be ready to go when you are. So whether you are video chatting, uploading family videos, or playing your favorite online games, all you need to do is click and you're there.



Download HD Movie ▾

Google Fiber is
200x faster

Google Fiber download time
7 seconds

Race Again

Your current download time
21 minutes 52 seconds



Group sessions

ca. 18:30	End of day 1	
19:00	Dinner at the Red Lion, Hinxton (hosted by ELIXIR)	
Day 2: Friday, 16 May		
09:00	Solutions for big data in other science communities	
	<ul style="list-style-type: none">• Earth satellite data	Wolfgang Lengert (ESA)
	<ul style="list-style-type: none">• Radio astronomy data	Arpad Szomoru (JIVE)
09:30	<i>Group sessions: Blue sky solutions for big data</i>	
10:30	Report from group sessions and discussion	
11:00	Break	
11:30	<i>Group sessions: Practicalities and actions needed to implement suggested solutions</i>	
12:30	Report from group sessions and discussion	
13:00	Lunch	
14:00	Closing discussion, roundup of challenges and solutions	Chair: Rafael Jimenez
15:00	End of workshop	

Closing discussion

ca. 18:30	End of day 1	
19:00	Dinner at the Red Lion, Hinxton (hosted by ELIXIR)	
Day 2: Friday, 16 May		
09:00	Solutions for big data in other science communities	
	<ul style="list-style-type: none">• Earth satellite data	Wolfgang Lengert (ESA)
	<ul style="list-style-type: none">• Radio astronomy data	Arpad Szomoru (JIVE)
09:30	<i>Group sessions: Blue sky solutions for big data</i>	
10:30	Report from group sessions and discussion	
11:00	Break	
11:30	<i>Group sessions: Practicalities and actions needed to implement suggested solutions</i>	
12:30	Report from group sessions and discussion	
13:00	Lunch	
14:00	Closing discussion, roundup of challenges and solutions	Chair: Rafael Jimenez
15:00	End of workshop	

Thanks to ...

- You!
- BMB



Stephanie Suhr

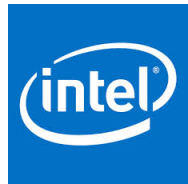
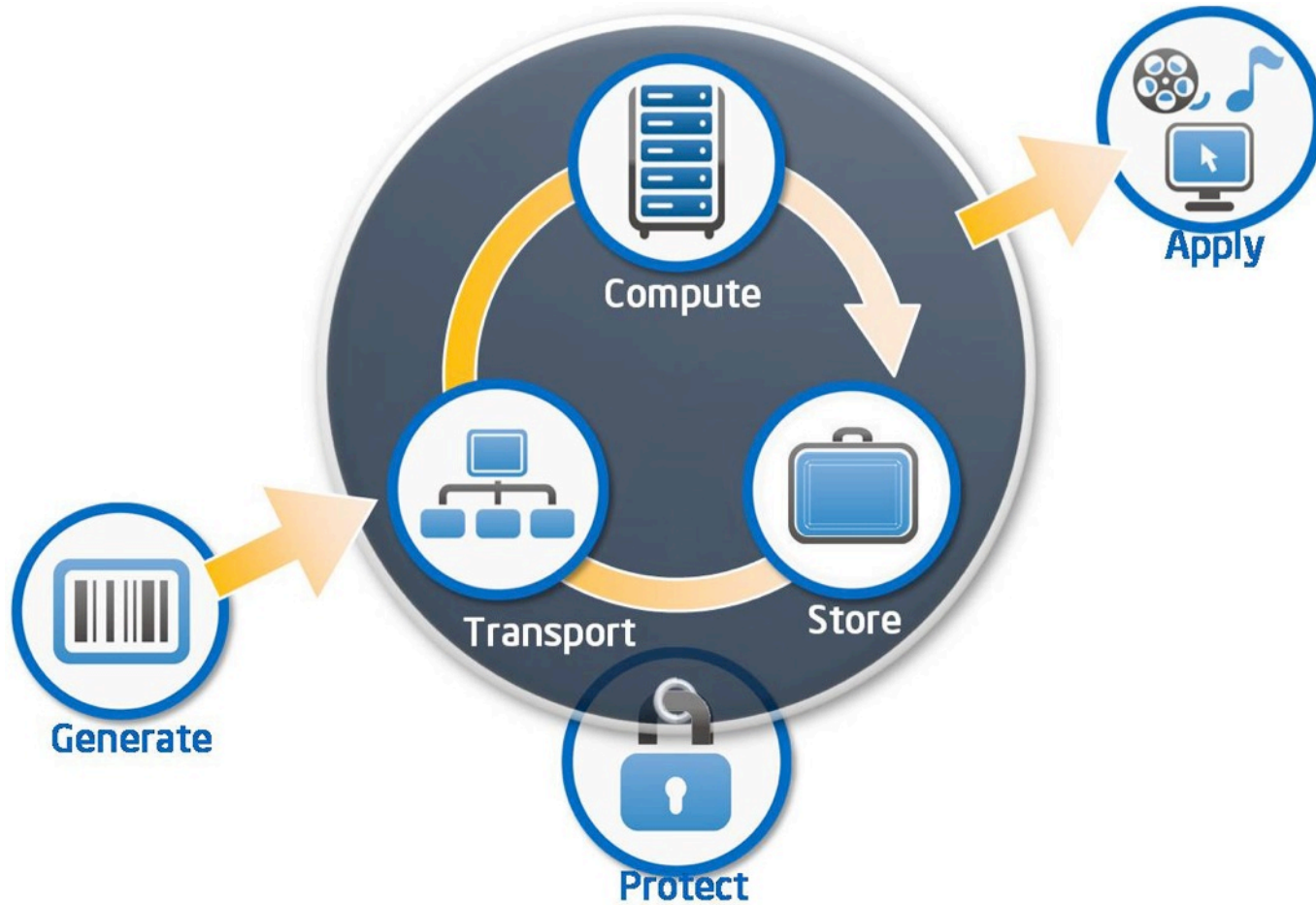


Tom Hancocks



Cath Brooksbank

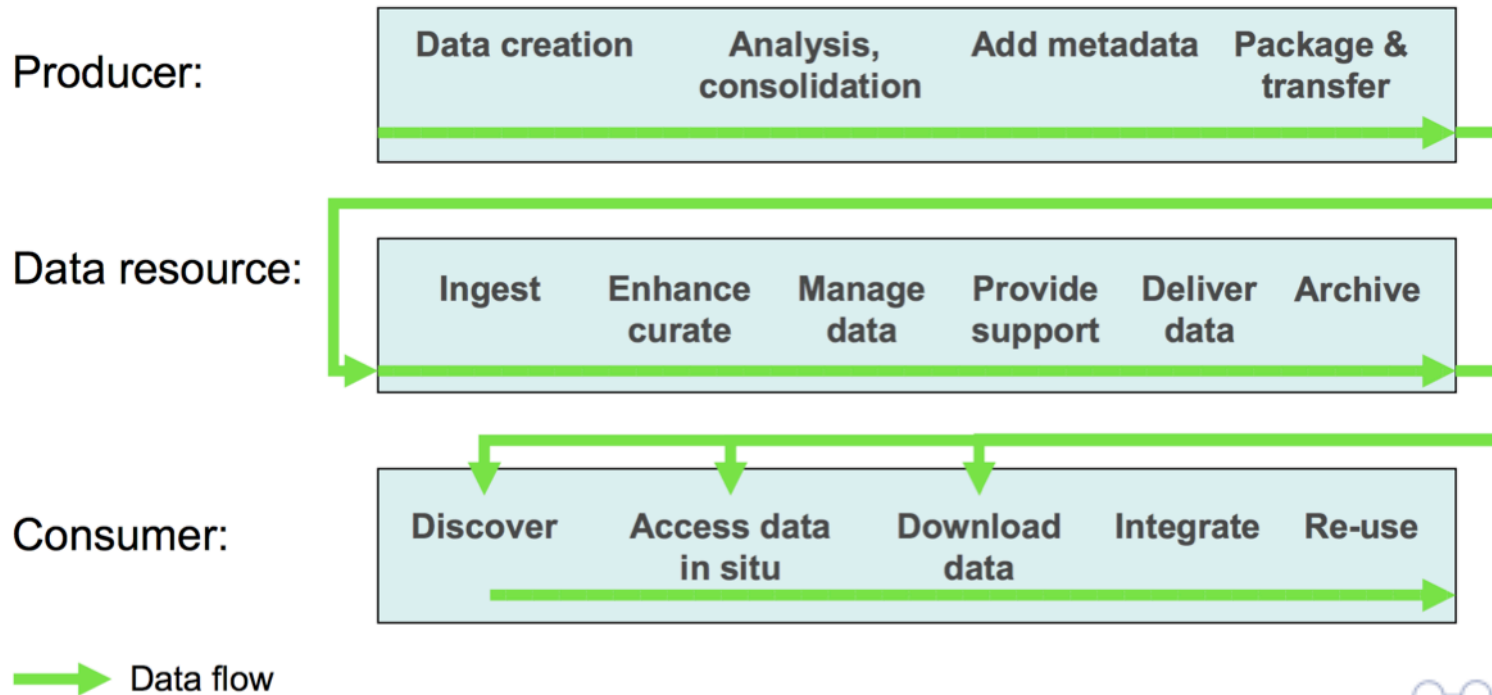
The Lifecycle of Data



<http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002552>

Large-scale data sharing in the life sciences

Processes and data flow



All of the
information

Information
you
need!

